From Dirty Data to Tidy Facts: Clustering Practices in Plant Phenomics and Business Cycle Analysis

Marcel Boumans and Sabina Leonelli

Abstract This chapter considers and compares the ways in which two types of data, economic observations and phenotypic data in plant science, are prepared for use as evidence for claims about phenomena such as business cycles and gene-environment interactions. We focus on what we call "cleaning by clustering" procedures, and investigate the principles underpinning this kind of cleaning. These cases illustrate the epistemic significance of preparing data for use as evidence in both the social and natural sciences. At the same time, the comparison points to differences and similarities between data cleaning practices, which are grounded in the characteristics of the objects of interests as well as the conceptual commitments, community standards and research tools used by economics and plant science towards producing and validating claims.

1 Introduction: Preparing Big Data for Analysis

Big data cannot be interpreted without extensive and laborious preparation, including various stages of processing and ordering to make it possible for data to be disseminated and subjected to analysis. Several chapters in this volume – including Halfmann's on sampling in oceanography, Karaca on data acquisition in particle physics and Hoeppe on sharing observations in astronomy - stress the decisive impact that such preparation practices have on the subsequent journeys of data and the use of data as evidence for claims about phenomena. In this chapter we discuss the epistemological significance of yet another practice of data preparation: *data cleaning*, that is the efforts involved in formatting, manipulating and visualising data so that they are sufficiently tractable to be amenable for analysis.

M. Boumans (✉)
Utrecht University School of Economics, Utrecht University, Utrecht, The Netherlands
e-mail: m.j.boumans@uu.nl

S. Leonelli
Department of Sociology, Philosophy and Anthropology & Exeter Centre for the Study of the Life Sciences (Egenis), University of Exeter, Exeter, UK

Alan Turing Institute, London, UK
e-mail: s.leonelli@exeter.ac.uk

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_5

The cleaning strategies that we aim to discuss are not focused on scrubbing and scraping dirt away, but rather on tidying up, sorting and ordering. In everyday life as in data practices, tidying up can be done in a variety of different ways depending on existing habits and future requirements. In what follows, we focus on two strategies for tidying up data which both rely, in different ways, on the clustering of objects into groups. The first strategy is to get rid of smudges and flecks by arranging objects so that unruly bits are less visible, and the eye is drawn to the more orderly – cleaner - parts of the ensemble. We exemplify this strategy through the analysis of data cleaning practices in economics, and specifically in relation to business cycle analysis, where data consist of observations of journalists, business annals, and social and economic statistical time-series. The second strategy is to put everything in boxes and store them some place out of sight, placing labels on each box to be able to retrieve its contents when needed (the more boxes and objects one has, of course, the more complex the labels will need to be).[1] We exemplify this strategy through the analysis of data cleaning practices in biology, and specifically the handling of phenomic data about plants, where data include images and measurements documenting the morphology, physiology and behaviour of organisms and their environments.

We compare a case from the natural sciences (biology) with one from social sciences (economics) in some detail to exemplify the complexity of the research practices involved, which mirrors the complexity of the phenomena under study in both areas. While the conceptual commitments, community standards and research tools used by economics and biology are starkly different, in both cases data cleaning and subsequent analysis involve bringing together voluminous datasets of diverse types and formats, generated by a broad range of heterogeneous sources. The projected value of these data as evidence for scientific claims grows with aggregation: the more data analysts are able to link together and consider as a single body of evidence, the more sophisticated and reliable the resulting insights are expected to be.

The chapter is organised as follows. In the first section, we examine the work required to create meaningful clusters from these forms of big data, and the extent to which data cleaning transforms datasets. In section two we draw on Mary Douglas's seminal analysis of dirt and impurity, in which she argued that cleaning is not about removal but about ordering, to identify a common strategy used by researchers in both cases, which we call *cleaning by clustering*. After discussing this general approach, we note how the specific mechanisms and tools used to enact this strategy differ considerably in the two domains of practice. In economics, cleaning by clustering is largely a question of exercising visual judgement grounded on principles similar to the Gestalt principles, thus arranging data in ways that are *aesthetically appealing and intuitively intelligible* to the analyst. This strategy goes a long way towards facilitating data mining, for instance through the construction of

[1] This approach to cleaning is heavily built on the strategies of packaging, curating and labelling explored by Leonelli (2011, 2016). Contrary to data packaging in her previous studies, however, tidying up is not primarily aimed at making data portable across contexts, but rather at making it possible for data to be analysed and interpreted.

data models that highlight meaningful correlations and direct analysts towards specific interpretations. By the same token, this form of clustering is difficult to undo, leading to a situation where the aesthetic criteria employed to arrange the data are traded off with the ways in which the data could be used as evidence. In plant phenomics, cleaning by clustering is instead guided by the attempt to define a "landscape" for the re-purposing of data: a set of conditions, in other words, through which researchers may be able to re-use data for new goals.[2] The priority in this case is not achieving visual intelligibility alone, but rather the creation of data visualisation and retrieval tools that enable users to *disaggregate data clusters when needed to confront new research questions*. This enables researchers to trace the origin of the relevant data journeys, and evaluate the reliability and appropriateness of every step of "cleaning" in light of novel situations of inquiry within which data may be re-purposed. We are particularly interested in identifying the principles that guide data cleaning activities in these cases, and the conceptual, material and social circumstances within which these principles are grounded and through which they originate. To this aim, in section three we explore the relation between data cleaning practices and how data are subsequently moved and used. Comparing our two cases points to significant differences between data practices, which are grounded in the nature of the objects of interest as well as in the conceptual commitments, community standards and research tools used by economics and plant science towards producing and validating claims. It also points to the difficulties experienced by data analysts in providing general principles of cleanliness with regard to research data, as exemplified by the recent debate around "tidy data" in computational data science, which we discuss in our closing section.

2 Cleaning Data: Empirical Cases from Plant Science and Economics

Our starting point is a close look at two cases of "data cleaning" taken from economics and plant science, respectively. The cases exemplify some of the most sophisticated forms of data processing in each field, aiming to encompass very different types and formats of data coming from a wide variety of sources, which can only be considered as a single body of evidence thanks to laborious processing. The economic case, concerning the generation of quantitative facts about the business cycle at the National Bureau of Economic Research in the 1940s, was selected for two reasons. On the one hand, this post-war research at the NBER is exemplary for many current practices of data preparation in economics, and on the other hand this practice was described so explicitly and in such great detail in a publication, *Measuring Business Cycle* (1946), that it enables and ensures insight and under-

[2] The landscape may include data collection strategies, repositories and visualisation tools enabling researchers to retrieve, compare and analyse data coming from a variety of sources.

standing of this specific clustering practice. The plant science case, concerning the processing of phenotypic data in plant phenomics, constitutes one of the most discussed examples of complex data processing in contemporary biology, with several ongoing debates documenting the rationale and strategies used to make data usable for further analysis. Below, we focus on the discussions surrounding the identification of essential data and related standards ("minimal information") for this kind of research.

2.1 Empirical Case: Measuring Business Cycles

Founded in 1920, the National Bureau of Economic Research (NBER) is a private, non-profit, non-partisan organization dedicated to conducting economic research and to disseminating research findings among academics, public policy makers, and business professionals.[3] The object of the NBER is "to ascertain and to present to the public important economic facts and their interpretation in a scientific and impartial manner" (Burns and Mitchell 1946, p. v). Wesley C. Mitchell, the first director of the NBER till 1945, was well-known for his contributions to the empirical analysis of business cycles.[4] The NBER is not a statistical office or bureau that aims at collecting economic and social data, but instead aims to analyse existing economic and social statistics, in this case to "measure business conditions." These statistics were data of various aspects of economic and business life and came from various different sources. An 11 page long appendix of *Measuring Business Cycle* (1946) list these statistics such as of industrial production, freight, sales, milk used in factory production, transit rides, railway passengers miles, wholesale prices, total income payments, employment, bank debits, electric power production, payrolls, business failures, from organisations such as Federal Reserve, Interstate Commerce Commission, Bureau of Foreign and Domestic Commerce, Railroad Companies, Bureau of Labor Statistics, Chicago Board of Trade, and Bureau of Foreign and Domestic Commerce.

The book *Measuring Business Cycles* (1946) was the result of 20 years of empirical business studies at the Bureau under the supervision of Mitchell. The aim was to identify and establish facts about the business cycles, which could be used to test existing business cycle theories. Burns and Mitchell stated that theoretical work on business cycles was "often highly suggestive; yet rest so much upon simplifying assumptions and is so imperfectly tested for conformity to experience that, for our purposes, the conclusions must serve mainly as hypotheses" (p. 4). At the same time, they observed that "satisfactory tests cannot be made unless hypotheses have been framed with an eye to testing, and unless, observations upon many economic

[3] See the NBER website, http://www.nber.org

[4] See Morgan 1990, pp. 44–56, for a more detailed background of the NBER and Mitchell's approach.

activities have been made in a uniform manner" (p. 4). Although theories were seen as "incomplete in coverage" and "highly suggestive," they were not "put aside" but used "as hypotheses concerning what activities and what relations among them are worth studying. In that way they will be of inestimable value in his factual inquiries" (p. 10). Hence the point of departure for data analysis was not a theory of the business cycle but a very general definition covering commonly accepted characteristics of the business cycles:

> Business cycles are a type of fluctuation found in the aggregate economic activity of nations that organize their work mainly in business enterprises: a cycle consists of expansions occurring at about the same time in many economic activities, followed by similarly general recessions, contractions, and revivals which merge into the expansion phase of the next cycle; this sequence of changes is recurrent but not periodic; in duration business cycles vary from more than one year to ten or twelve years; they are not divisible into shorter cycles of similar character with amplitudes approximating their own. (Burns and Mitchell 1946, p. 3)

This working definition was supposed to list the observable characteristics of a "distinct species of economic phenomena" (p. 3), that is the business cycle. This definition focused on what should be measured, such as the average duration of the cycle. To achieve this aim, all kinds of questions raised by this definition had first to be answered.[5]

To understand which principles of clustering were used in this case of business cycle measurement, we need to have a closer look at the four implicit assumptions made within this definition. The first assumption is that the cyclical turns of different processes are concentrated around certain points in time. The second assumption is that the business cycle is not a periodic but a recurrent process, a "regularity" that is different from "seasonal variations, random change, and secular trends" (p. 6). Another assumption of the definition is that business cycles run in a continuous round, "no intervals are admitted between one phase and its successor, or between the end of one cycle and the beginning of the next" (p. 7). And the last assumption is the duration of the cycle, somewhere between 1 year and 10 or 12 years.

The main problem for analysts is that business indexes and time series do not show "cyclical patterns" that are "sweeping smoothly upward from depressions to a single peak of prosperity and the declining steadily to a new trough" (p. 7), and so a business cycle has to be identified from an irregular process, where the movements are interrupted by others in the opposite direction, and where one may see double or triple peaks and troughs. What therefore is needed are criteria to identify the characteristics of the business cycles, such as "what reversals in direction mark the end of a cyclical phase" (p. 8). Crucial to our analysis is the fact that such criteria cannot be derived from any (business cycle) theory,[6] but rather they relate to aesthetic

[5] Such as, for instance: How large or small does a nation have to be to have a business cycle, or is it an international phenomenon? How far back in time can business cycles be traced? What is the most appropriate level of aggregation? Which economic activities should be included?

[6] See Bogen and Woodward (1988) for a similar, more general claim about the incompleteness of theories in this respect.

judgements based on visual displays of the data. In other words, certain smooth and simple shapes turn out to be used as tools to process and visualise the data. The approach is based on pattern recognition, described by Burns and Mitchell (1946, p. 8n) as "the source of all true knowledge", but nevertheless it is required to be as objective as possible. Indeed, these criteria are presented as "a 'brake' on an investigator's pattern sense which […] may lead to mischievous fictions" (p. 8n).

Burns and Mitchell emphasized that the cyclical pattern can be seen "only by the eye of the mind" (p. 12). "What we literally observe is not a congeries of economic activities rising and falling in unison, but changes in readings taken from many recording instruments of varying reliability" (p. 14). To "see" the business cycle "in the mind's eye," these recordings have "to be decomposed for our purposes; then one set of components must be put together in a new fashion" (p. 14).

> We conceive business cycles to consist of roughly synchronous movements in many activities. To determine whether this thought symbol represents experience or fantasy, our measures of the cyclical behavior characteristics of many activities must be assembled into the end products of which our definition is the blueprint. In statistical jargon, time-series analysis must be followed by a time-series synthesis. (Burns and Mitchell 1946, p. 17)

The idea is the decomposition of the time series into cyclical, secular, seasonal and random movements, but the "isolation of cyclical fluctuations" was considered to be a "highly uncertain operation" (p. 37), particularly if it is done in a "mechanical manner". The components cannot be segregated without considerable testing and experimenting by skilled technicians. "There is always danger that the statistical operations performed on the original data may lead an investigator to bury real problems and worry about false ones" (p. 38).[7]

Most of the analysis was in the determination of cyclical timing. It had become clear that the data needed to be adjusted for – i.e., cleaned from – seasonal variations "to be more useful in explaining business cycles than would measures made from highly fabricated data" (p. 43). We therefore briefly focus on this aspect of the business cycle analysis, to show how much it was a combination of "hunch and judgment" (p. 44) and mechanical methods, which results were evaluated based on their visual displays.

Two methods were used, one consisted in taking averages of the original figures for each months, which were adjusted for secular trend; and the other entailed taking a 12-month moving average of the original figures, placing each average in the seventh month of shifting 12-month intervals. The rationale for both methods are the assumptions that "random components of a series [will] cancel one another" and that "the process of averaging will tend also to make the cyclical component of a series sum to zero" (p. 47).

When the data was adjusted for seasonal variations, the next problem was the dating of cyclical fluctuations. Therefore the data was plotted upon a semi-logarith-

[7] See Boumans 2015 for a more detailed account of measurement, which sees measurement as a considered balance between mechanical objectivity and expert judgement.

mic chart (typically about 7 feet long) such that the whole record was studied in this graphic form. As far as possible the scales were kept uniform.

The basic criterion for distinguishing the three types of movements, that is the cyclical, secular and erratic movements, was their duration. Secular trends were conceived as drifts that persist in a given direction for a few decades. Erratic movements, the "saw-tooth contour" (p. 57) were supposed to cover no longer than a few months. But even with this basic criterion, the judgments were often difficult:

> When specific cycles are made doubtful by random movements, we smooth the data by moving averages and base judgments upon the curve of moving averages. When the secular trend rises sharply, we allow brief and mild declines to count as contractions of specific cycles. Similarly, when the secular trend falls sharply, brief and mild rises are counted a specific-cycle expansions. (Burns and Mitchell 1946, p. 57)

Once the cycles had been distinguished the NBER researchers proceeded with the dating of the turning points. The idea is to take the highest and lowest points of the plotted curves as the dates of the cyclical turns. But often it is not clear to decide which points these are, for example when erratic movements are prominent in the vicinity of a cyclical turn. Then all kinds of checks or averages have to be considered to arrive at a determination.

> Our methods of determining specific cycles make no pretensions to elegance. Since no fast line separates erratic or episodic movements from specific cycles, or erratic turns from cyclical turns, there is ample opportunity for vagaries of judgment. At times our rules fail to yield a clear-cut decision. At times the members of our statistical staff disagree in their efforts to apply the rules to a given series. Our experience indicates that this difficulty cannot be removed by multiplying rules. (Burns and Mitchell 1946, p. 64)

The judgment is instead based on a consensus of three persons who have worked independently on marking off the cycle. Once arrived at this consensus, the whole process is audited by an "experienced member of the staff" (p. 64) (Fig. 1).

2.2 Empirical Case: Processing and Interoperability Requirements for Imaging Data in Plant Phenomics

Plant phenotyping involves analysing plant trait data with the aim to study development and gene-environment interactions. It emerged in the 1960s with an initial emphasis on quantitative analysis, which was later broadened to imaging data obtained via high-throughput experiments performed in fields, glasshouses, and/or laboratories. Such imaging data, and the accompanying observations about the conditions under which the images were obtained, now constitute the most coveted type of data in this field, with increasingly sophisticated tools being developed for their visualisation and automated analysis. This shift of emphasis on complex data formats proceeded in parallel to the broadening of the term "phenotyping" to include any type of morphological variability within organisms, thus encompassing not only the immediately visible features of organisms, but also (1) features of tissues,

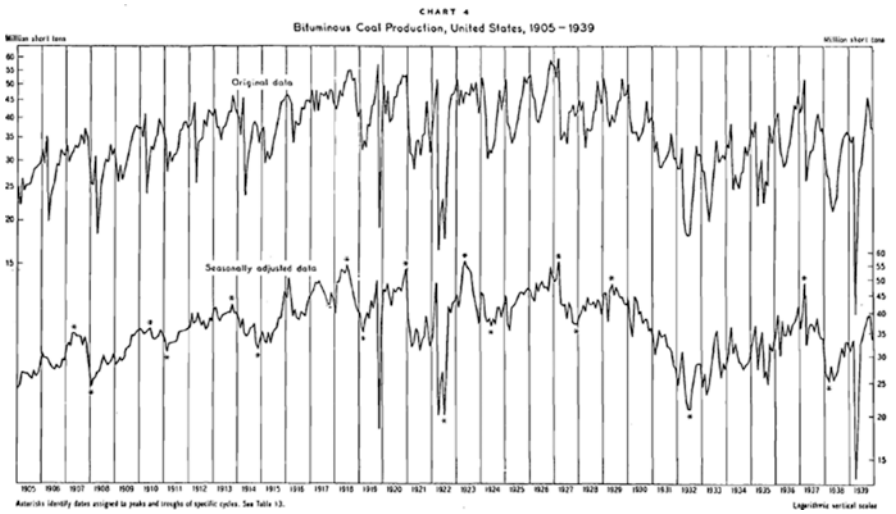


Fig. 1 Example chart of a time series in its original shape and after it has been adjusted for seasonal variation. The adjustment is supposed to facilitate dating of turning points, indicated by the asterisks. (Source: Burns and Mitchell 1946, p. 60, Chart 4)

proteins, metabolic pathways and other aspects only accessible through intervention and specialised imaging techniques; and (2) the ways in which such features vary across environments that range from laboratories to glasshouses, field trials and the "wild" – which involves collecting data on the soil, climate, other organisms and microbiome with which plants interact. In the words of prominent contributors to the field, phenotyping – also called "phenomics" – "broadened its focus from the initial characterization of single-plant traits in controlled conditions towards 'real-life' applications of robust field techniques in plant plots and canopies" (Walter et al. 2015). Importantly for our analysis, this shift in the conceptualisation of phenotypic traits made them much less obviously identifiable as concrete descriptors. Collecting data about the size of a leaf or the structure of a metabolic pathway is not simply a matter of observation, but rather is informed by a rich conceptual apparatus defining what counts as leaf surface and metabolism. Thus, just as much as business cycles are no pure theoretical constructs, phenotypes are no 'brute facts' about the world: in both cases, empirical and theoretical considerations remain firmly intertwined, and affect researchers' approach to data processing and interpretation.

A key component of contemporary phenomics, and the reason why it is regarded as generating knowledge that can underpin and guide agricultural production, is a holistic characterisation of plant performance, which involves the employment of several investigative methods and the generation and analysis of a wide variety of data types. These include, for instance, multispectral and thermographic imaging of plant growth, which is often carried out within so-called "smart glasshouses" in an automated fashion (by robots or conveyor belts that transport the plants to various imaging chambers, multiple times per day, over an extended period of time).

Photographs and measurements are produced that document how plants develop, how their leaves and roots change, and how they respond to external stimuli.

Cleaning such images for analysis involves judgements around the quality and resolution of the photograph, the lighting and background conditions, the position in which plants have been captured and the extent and clarity to which relevant leaves and roots show in the picture. The quantity of images generated through any one experiment makes it hard for researchers to do such work manually, and yet it is hard to fully automate due to the large amount of know-how and theoretical commitments involved in judging image quality – encompassing familiarity with the plants and their full life-cycle, expectations around how plants may respond to environmental conditions, existing conceptualisations of plant development and growth, and assumptions around which environmental and morphological elements need to be valued and prioritised over others.

Another popular type of phenomic data is acquired through top-view imaging of the plant canopy in the field, which can be performed by humans in helicopters, robots or remote-controlled drones. These photographs can be analysed to measure leaf greenness, via tools such as the Normalised Difference Vegetation Index, or plant biomass and growth in the area under scrutiny. Again, while some basic parameters can be established for what counts as a "bad image" and which elements of each image may be classified as "noise", cleaning such images involves expert assessment based on detailed knowledge of the characteristics and patterns of growth of the plants at hand. An example (Fig. 2) is an imaging study of soy-bean fields to determine patterns of growth, in which researchers prepare images for further analysis (in their own words, "classify" the images) through models that are manually trained at every step to respond to the traits of interest in the beans (Xavier et al. 2017).

Given the sensitivity of phenomic studies to local conditions and the conceptual preferences and know-how of specific researchers, consensus around how to clean data is hard to achieve. Nevertheless, such consensus is highly valued and sought for, as it enables researchers to compare results obtained across species, field types and environmental conditions. One attempt towards establishing general standards for data collection and processing is the Minimal Information About Plant Phenotypic Experiments, or MIAPPE. MIAPPE is part of a broader set of "minimal information about data" movement now recognized and coordinated by the FAIR sharing international initiative for reusable data curation.[8] This is an attempt to standardize the practices and variables required to tidy up data formatting and analysis enough to make data searchable, visualisable and retrievable through digital means. The idea of "minimal" information is meant to foster an evaluation of which contextual information is most important to data interpretation, resulting in as small a set

[8] See https://fairsharing.org/collection/MIBBI. Among the first incarnations of the movement, and now highly successful standards in their own right, were the Minimal Information About a Microarray Experiment, or MIAME (Rogers and Cambrosio 2007) and the Minimal Information for Biological and Biomedical Investigations, or MIBBI (http://www.nature.com/nbt/journal/v26/n8/full/nbt.1411.html)

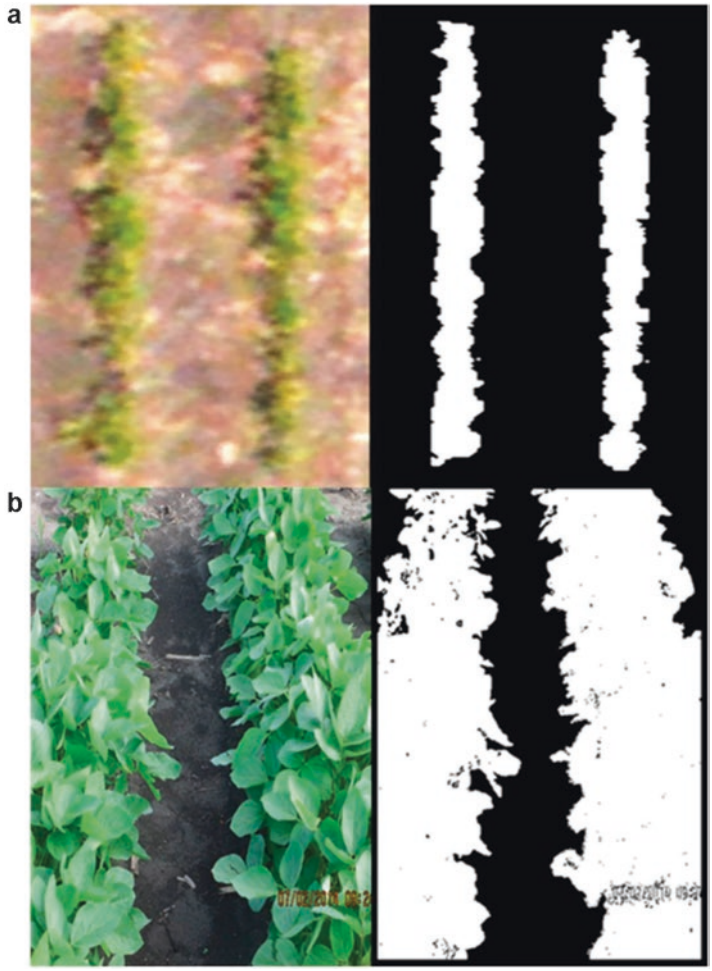


Fig. 2 Example imagery of a single plot of soy-bean canopy, used to calculate a percentage canopy coverage on a given sampling date. (**a**, **b**) From aerial (above; **a**) or ground (below; **b**) platforms, with raw (left) and classified (right) imagery. (Source: http://www.genetics.org/content/206/2/1081)

as possible of metadata that researchers view as essential to phenotypic data reuse. Somewhat paradoxically, within MIAPPE this aspiration towards minimal information is accompanied by the wish to lose as little information as possible about the original format of the data, the circumstances under which they were generated, and the ways in which they were processed since. This is because the specificity of the provenance and formatting of data in each case is regarded as highly valuable by the plant scientists using such data for their own research, a requirement that researchers and engineers involved in the development of MIAPPE take seriously: "We had to allow for differences that occur between particular types of plant experiments, e.g. performed in different growth facilities. This is reflected in a varying set of attributes recommended in MIAPPE" (Ćwiek-Kupczyńska et al. 2016). Indeed, the

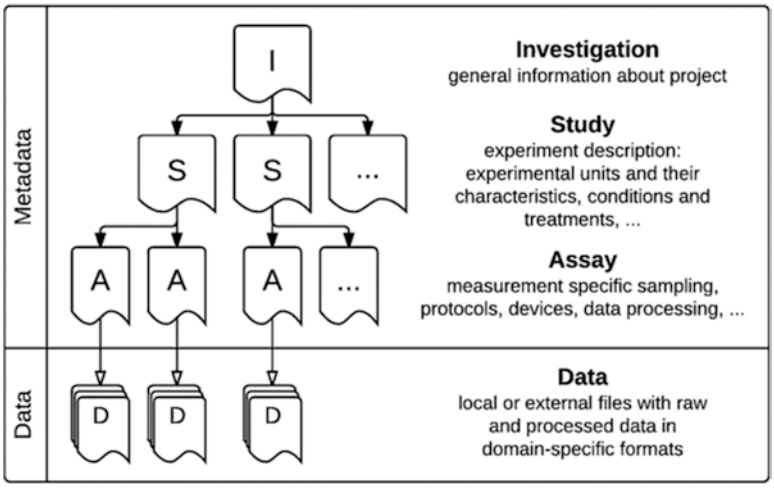
list of attributes to be reported to MIAPPE involves over 80 items, which can extend to over a hundred depending on the field conditions. The basic categories are themselves relatively broad, encompassing general metadata, timing and location, biosources, environment, treatments, experimental design, observed variables and as much information as possible on sample collection, processing and management – a far cry from the minimalism that the MIAPPE criteria were expected to exemplify.

It is useful to consider a couple of the simplest examples from this list. Take for instance the item "location and timing of an experiment". Here MIAPPE developers note that "depending on the nature of the study and scientific objectives, different initial time points might be crucial—sowing date or transfer date, treatment application time, etc. The duration of particular stages is also important." (Ćwiek-Kupczyńska et al. 2016, p. 3). Thus, even a relatively straightforward measure such as the time of the experiment turns out to be a complex and context-dependent issue, for which it is hard to establish any hard and fast boundaries to ensure comparability across different experiments.[9] Another example is item "biosource" – that is, the identification of the plant material at hand. Here MIAPPE recommends using at least two attributes, one consisting of the species name as in standard taxonomic classifications, and the other consisting of the "infraspecific" name, pointing to the specific variant, accession or line in question. Complications arise due to the types and history of the plant materials at hand. While the taxonomy of plant species is, though controversial, subject to international standards, the identification and classification of sub-species variants is highly decentralised and context-dependent, with no overarching agreement around classification and often not even a clear awareness of the differences between local systems. For example the varieties of the plant *Manihot esculenta*, whose root cassava is a key crop in West Africa and South America, are often defined by the different ways in which local breeders value specific traits (like the humidity and colour of the root) when processing the plant for food production. Aware of this fact, the authors point to the importance of referencing any "public collection of names", and/or a specific experimental station or genebank in which the variant may be stored and or the seeds may have been sourced, and to which they can be physically traced. There are international identification systems for crops of commercial interest, such as the FAO/Bioversity Multi-Crop Passport Descriptors, but these do not cover all possible variants. The ways in which data about specific attributes are structured in MIAPPE conform to the ISA-Tab standards for data ordering, which is widely adopted in biology and looks as follows (Table 1).

This table aims to impose a clear conceptual ordering of the data, resulting in their presentation in a format and structure that is amenable to computational analysis. At the same time, the application of the ISA-Tab standard to the specific case of phenotyping is complex, as demonstrated by challenges encountered in developing the so-called "ISA-Tab Phenotyping Configuration". This consists of a standard Investigation file, a Phenotyping Assay file describing phenotypic procedures and observed variables (according to the dozens of attributes identified by MIAPPE,

[9] See Leonelli (2018) for an analysis of data time and its significance particularly within experiments.

Table 1 The structure of an ISA-Tab dataset



Source: Ćwiek-Kupczyńska et al. (2016)

such as location and biosources), and three versions of a Study file: one called "basic study" and consisting of a default general description of all plant experiments, which needs to be extended by added recommended MIAPPE attributes as applicable to the specific case[10]; and two extensions called "field" and "greenhouse" studies, featuring specific attributes for growth facilities and environmental information (Ćwiek-Kupczyńska et al. 2016, p. 8) (Table 2).

Notably, despite the drive towards comparability, MIAPPE emphasizes the need to capture any data format in use within the relevant scientific communities, rather than attempting to impose overarching standards on the ways in which data are produced: "in our implementation of MIAPPE, we do not restrict the format of the raw data in any way; it can be any custom, platform- or device- specific format, including texts, images, binary data, etc." (Ćwiek-Kupczyńska et al. 2016, p. 11). At the same time, MIAPPE requires that information about data provenance (metadata) is reported in ways that are comprehensive and retrievable by later data users. The most stringent MIAPPE instructions concern how to organize and display such metadata:

> If there is no description, the Derived Data File should be a standard, plain tab-separated sample-by-variable matrix. Its first column should contain (in the simplest situation) values from the Assay Name column in the Assay file, and the rest of the columns provide values for all variables. The names of those columns should correspond to the values in the Variable

[10] In practice, it can be also used when very little is known about the origin of observations, e.g. for simple, external or legacy phenotypic datasets that should be formatted as ISA-Tab, without the ambition to satisfy the MIAPPE recommendations.

Table 2 Illustration of what the basic ISA-TAB fields correspond to when implemented by plant scientists in the field and in the greenhouse, respectively

Basic	Field	Greenhouse
Source name	Source name	Source name
Characteristics[organism]	Characteristics[organism]	Characteristics[organism]
Characteristics[Infraspecific name]	Characteristics[Infraspecific name]	Characteristics[Infraspecific name]
Characteristics[seed origin]	Characteristics[seed origin]	Characteristics[seed origin]
Characteristics[study start]	Characteristics[study start]	Characteristics[study start]
Characteristics[study duration]	Characteristics[study duration	Characteristics[study duration
Characteristics[growth facility]	Characteristics[growth facility]	Characteristics[growth facility]
Characteristics[geographic location]	Characteristics[geographic location]	Characteristics[geographic location]
	Protocol REF[rooting]	Protocol REF[rooting]
	Parameter value[rooting medium]	Parameter value[rooting medium]
		Parameter value[container type]
		Parameter value[container volume]
	Parameter value[plot size]	Parameter value[container dimension]
	Unit	Unit
	Parameter value[sowing density]	Parameter value[number of plants per container]
	Parameter value[pH]	Parameter value[pH]
	Protocol REF[aerial conditions]	Protocol REF[aerial conditions]
	Parameter value[air humidity]	Parameter value[air humidity]
	Parameter value[daily photon flux]	Parameter value[daily photon flux]
	Parameter value[length of light period]	Parameter value[length of light period]
	Parameter value[day temperature]	Parameter value[day temperature]
	Parameter value[night temperature]	Parameter value[night temperature]
	Protocol REF[nutrition]	Protocol REF[nutrition]
	Parameter value[N before fertilisation]	Parameter value[N before fertilisation]
	Parameter value[type of fertiliser]	Parameter value[type of fertiliser]
	Parameter value[amount of fertiliser]	Parameter value[amount of fertiliser]

(continued)

Table 2 (continued)

Basic	Field	Greenhouse
	Protocol REF[watering]	Protocol REF[watering]
	Parameter value[irrigation type]	Parameter value[irrigation type]
	Parameter value[volume]	Parameter value[volume]
	Parameter value[frequency]	Parameter value[frequency]
	Protocol REF[sampling]	Protocol REF[sampling]
	Parameter value[experimental unit]	Parameter value[experimental unit]
Sample name	Sample name	Sample name

Source: Ćwiek-Kupczyńska et al. (2016)

> ID column in the Trait Definition File […]. So, a default derived data format is an "Assay Name × Variable" matrix of observations, that can be quantitative or qualitative. An extension of the above rule governing the format of the Derived Data File is possible by using values from another "data node" column (e.g. Source Name, Sample Name, Extract Name, etc.) as unique identifiers of the rows in the table with the associated observations. (Ćwiek-Kupczyńska et al. 2016, p. 12)

This is because such ordering is what enables researchers to initiate comparisons:

> we can provide separate data files with measurements taken for different observational units, e.g., morphological traits like "height" and "number of leaves" can be assigned to the whole plant, whereas physiological traits can be restricted to samples taken from particular leaf of a plant. Also conveying data aggregated over "data nodes" is possible in this way. (Ćwiek-Kupczyńska et al. 2016, p. 12)

Despite the attention placed by MIAPPE developers on the variability and contextuality of data and related preparation procedures, applying MIAPPE criteria to the processing of data in the field remains a big challenge. As a concrete example, we take the data processing performed at a leading station for the collection of phenomics data in the UK. The North Wyke Farm Platform is a research facility built around a working farm in Devon, in which researchers can study the interactions between climate, soil, animals, plants and microbiota in as close a setting as possible to real farming. The whole area is full of sensors and measurement devices, which collect data at regular intervals (15 minutes) about a variety of aspects of the farm: temperature, soil composition, humidity and rainfall, etc. The sensors are calibrated and checked in 15 huts ("monitoring cabins") positioned around the fields, and the data produced is sent wirelessly to the central computing facility based in the manor house, where researchers proceed to prepare the data, cluster them and store/disseminate them through a database. There are also three meteorological stations that move around the fields. An important activity besides collecting numerical measurements is the collection of samples (of soil, air, water, insects and plants),

which are acquired manually (e.g. manual sampling device for soil), prepared and stored in fridges at various temperatures).[11]

Researchers interviewed[12] in North Wyke have stressed that the data collected by the Farm Platform are not yet being interpreted: this will only be possible when enough longitudinal data are collected over the course of the next few years.[13] This makes the task of data cleaning ever more important, since the researchers' main task at the moment is to make sure that the data collected is reliable and clustered and displayed in ways that will facilitate further analysis, and prove informative for interested farmers. Cleaning the data means first of all making them comparable and consistent with other datasets generated within the Farm, an arduous task given the variety of measurements taken and images collected. Equally important is to make sure that such data would be comparable and consistent with other phenomics data from outside North Wyke. While researchers attempt to follow criteria similar to those formulated by MIAPPE, the variability in the interpretation of the attributes and values is a serious threat to automated mining and comparison among the data. Researchers aim to enable analysis in the future, but caution against any automated search. They also emphasize how the power of this evidence is in the meta-data, the information that enables researchers to contextualize the findings and evaluate their significance in relation to findings from other locations enacting different epistemic cultures and methods.

3 Cleaning by Clustering: The Principles Underpinning Data Cleaning Practices

Renowned anthropologist Mary Douglas provided an important argument for understanding the process of cleaning as being not about removal, but about ordering. According to Douglas (2002), dirt is essentially disorder: "There is no such thing as absolute dirt: it exists in the eye of the beholder. […] Dirt offends against order. Eliminating it is not a negative movement, but a positive effort to organize the envi-

[11] The facility attracts researchers from different communities and disciplines seeking to develop sustainable agriculture and ruminant production systems http://www.nature.com/news/agriculture-steps-to-sustainable-livestock-1.14796. It is the only currently functioning facility of its kind world-wide, and the Global Farm Platform http://www.globalfarmplatform.org/ was born to attempt to export this model and initiate similar sites elsewhere.

[12] Interviews were carried out by Leonelli in January 2016. A subset of the interviews, which interviewees consented to release in an open access format, is available here: https://zenodo.org/communities/datastudies/?page=1&size=20

[13] North Wyke researchers are also conducting short-term studies in which the data are used as evidence for claims about phenomena. Examples include research on replacing nitrogen as fertilizer, the use of plants to manage soil and water during floods, shifts in soil biota as land use changes, and the modelling of grassland production systems. At the same time, researchers only take up research that will not "distort" on-going, long-term data collection by forcing them to "clean" data with too narrow a set of epistemic goals in mind.

ronment" (p. 2). In chasing dirt when tidying we are "positively re-ordering our environment, making it conform to an idea […] it is a creative moment, an attempt to relate form to function, to make unity of experience" (p. 3). Douglas emphasizes that the identification of dirt should not be considered as a unique, isolated event. "Where there is dirt there is system. Dirt is the by-product of a systematic ordering and classification of matter, in so far as ordering involves rejecting inappropriate elements" (p. 44). Cleaning is the reaction which condemns any object or idea likely to confuse or contradict cherished classifications, thus "reducing dissonance" (Douglas 2002, p. 340). Thus cleaning is part of the epistemological activity of systematization, such as ordering and classification. Douglas distinguishes two phases to such systematization practices:

> In the course of any imposing of order, the attitude to rejecting bits and pieces of dirt goes through two stages. First they are recognisably out of place, a threat to good order, and so are regarded as objectionable and vigorously brushed away. At this stage they have some identity: they can be seen to be unwanted bits of whatever it was they came from, hair or food or wrappings. This is the stage at which they are dangerous; their half-identity still clings to them and the clarity of the scene in which they obtrude is impaired by their presence. But a long process of pulverizing, dissolving and rotting awaits any physical things that have been recognized as dirt. In the end, all identity is gone. The origin of the various bits and pieces is lost and they have entered into the mass of common rubbish. It is unpleasant to poke about in the refuse to try to recover anything, for this revives identity. So long as identity is absent, rubbish is not dangerous. It does not even create ambiguous perceptions since it clearly belongs in a defined place, a rubbish heap of one kind or another. (Douglas 2002, pp. 197-8)

The stage of total disintegration is the stage in which dirt has become undifferentiated. Then a cycle has been completed, resulting in an order that is either continuous with what was there before the cleaning or created by the process of cleaning itself.

Drawing on Douglas's analysis, we argue that in both of our cases researchers adopt the same broad strategy for data cleaning: they *clean by clustering*. Cleaning is a way to impose order and intelligibility on a dataset, by identifying categories and typologies for classification, models and algorithms through which data can be filtered and selected, and/or tools through which data can be displayed and organised so as to enable further analysis and interpretation.

The specific mechanisms and tools used to enact this strategy, however, differ considerably across our cases, revealing a divergence in the heuristic principles used to guide and motivate the cleaning strategies, and the extent to which whatever is neutralized from a given stage of data cleaning is regarded as "unwanted bits" with "some half-identity clinging to them", or as dirt where "identity is absent".

In our economics case, clustering involves looking for cyclical patterns through visual judgement. To understand the heuristic behind this cleaning procedure, it is useful to discuss briefly Gestalt theory first. Gestalt psychologists study perceptual organization: "how all the bits and pieces of visual information are structured into larger units of perceived objects and their interrelations" (Palmer 1999, p. 255). A "naïve realist" explanation of this organization could be that this organization simply reflects the structure of the external world. A problem with this explanation is that the visual system does not have direct access to how the environment is structured, it has only access to the image projected onto the retina, the "array of light that falls on the

retinal mosaic" (p. 257). This optic array allows for an infinite variety of possible organizations. The question therefore is how the visual system picks out one of them. To answer this question Max Wertheimer, one of the founders of Gestalt psychology, studied the stimulus factors that affect perceptual grouping: "how various elements in a complex display are perceived as 'going together' in one's perceptual experience" (Palmer 1999, p. 257). The theoretical approach of the Gestalt psychologists is that perceptual organization is grounded in the wish to maximize simplicity, or equivalently, minimize complexity. They called this hypothesis the principle of Prägnanz, today also called the minimum principle. It states that the percept will be as good as the prevailing conditions allow. The term "good" refer to the degree of figural simplicity or regularity, and the prevailing conditions refer to the structure of the current stimulus image (Palmer 1999, p. 289). The Gestalt psychologists saw symmetry as a global property with which figural goodness could be analysed.

The organising Gestalt in the case of the NBER business cycle analysis was a cyclical pattern, such as the Fig. 3. By taking averages, whether weighted or not (which is an act of clustering), one aimed at reducing the noise in the observations as much as possible. Because it is not possible to tidy up by a kind of physical intervention on some physical material, the tidying up is not done by removal but by clustering in such a way that the cluster itself is "cleaner" than the individual data. The principle of Prägnanz that was implicitly applied and was the underlying goal of the procedures is an as simple as possible shaped cycle with clear peaks and troughs.

In the economic case, the original data end up as what Douglas classified as undifferentiated dirt – that is, as objects that are forever disconnected from their original source.

> [T]hese symbols are derived by extensive technical operations from symbolic records kept for practical ends, or combinations of such records. We are, in truth, transmuting actual experience in the workaday world into something new and strange […]. (Burns and Mitchell 1946, p. 17)

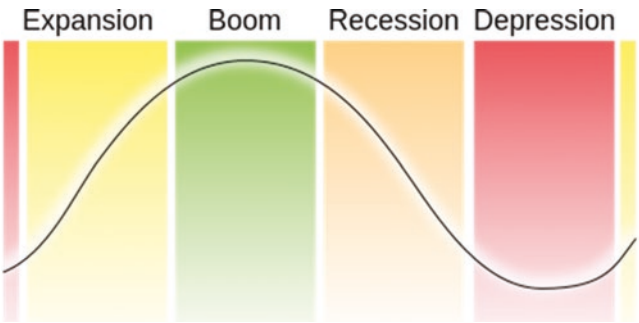


Fig. 3 Example of a "typical" business cycle pattern (Source: https://seekingalpha.com/article/2716385-investing-in-business-cycles)

In other words, the process of cleaning by clustering in this case transforms a large quantity of objects that were previously identified as data into objects that have new evidential value, but are no longer available or retrievable as sources of information about the contexts from which they were inferred.[14] At the same time, it is important to note that the resulting records do not completely fail to provide an identity to the discarded objects. Keeping some traces of the original time series is relevant if only to verify that results are not artificial products of spurious cyclical patterns. The visualisations of original times and the adjusted one should show sufficient similarity. "A common method of judging the goodness of [an] adjustment is to see whether the adjusted figures show similar movements in successive years" (Burns and Mitchell 1946, p. 54).

In plant phenomics, clustering instead involves defining a "landscape" for the potential re-contextualisation of data. The starting assumption is that phenomics data, in all their richness, variability and multiplicity of features, may be used for all sorts of research goals, ranging from studies of irrigation systems to investigations of plant growth and nutrition (as in the case of North Wyke data). Therefore the priority for researchers is not the visual intelligibility of a particular way of arranging data, but rather the creation of categorisations that facilitate the *disaggregation* of data clusters when needed by the inquiry at hand. In other words, researchers want to retain the ability to trace the origin of the relevant data journeys, and evaluate the adequacy of every step of data cleaning towards producing reliable evidence for new research questions. Key heuristic principles here are: *accuracy*, in the sense of being as faithful as possible to the specific characteristics of the research objects at hand; and *traceability* of data sources, in the sense of making sure that prospective data analysts have what they need to assess the quality of the data and, if needed, process them differently (which typically includes as extensive an access as possible to metadata).

This approach is hard to compare to the application of Gestalt principles, because those are focusing on visual appearance and presentation, while phenomics practices of cleaning by clustering focus on interpretability and the potential to disaggregate existing data clusters. Nevertheless, like the economics case, this is in striking opposition to common sense interpretations of the metaphors of "cleaning" and "dirt" that focus on the removal of blatantly unwanted items. Both in biology and economics "dirt" may (and often does) contain useful information, which needs to be ordered so as to be retrievable depending on the interests of the prospective analyst. The original datasets and related metadata never fully become undifferentiated dirt as in Douglas's analysis. Rather, researchers attempt to "cling on to their half-identity", in Douglas's terms, thus leaving open the option for these objects to be re-identified as data and fully reinstated as significant sources of evidence for a claim. The main difference between the two fields is that economic data have lost more of the identity of their original data than is the case in phenomics. While in plant phenomics accuracy and traceability are leading, in economics accuracy has to be balanced with Prägnanz, and traceability is not required.

[14] This interpretation assumes a relational account of data epistemology, as outlined in Leonelli (2016) and in the introduction to this volume.

4 Comparing Heuristics across Research Communities in Natural and Social Sciences

Economic data are processed in ways that make them much more computationally tractable than phenomics data due to their numerical format. Economic data are thus better amenable to aggregation and analysis in comparison to many other data types, which potentially expands their scope for linkage and aggregation with other datasets but also limits the power of investigators to contextualise and situate the data in relation to their origin. In this case, cleaning by clustering is a cumulative process, in which the bulk of "raw" data is replaced by a smaller set of business-cycle "facts" through the exercise of visual principles.[15] As a result, analysts working at later stages of these data journeys are left mostly with data models that conform to specific criteria and are best used to address a narrow set of questions, in conformity with the principles and assumptions made while preparing them for analysis. The original "raw" data are no longer accessible, having been "cleaned out" in the data visualisations.

By contrast, phenomics data remain more difficult to analyse through computational tools, and can only be compared and linked with other datasets by employing case-by-case adjustments. They are so heterogeneous, and their ordering into clusters so pluralistic and open to multiple interpretations, that additional processing is needed every time researchers re-use them for a specific project. When considering data on biosource as discussed in section two, for instance, researchers need to double-check what assumptions have been made about the taxonomy of plant varieties when ordering plant traits into groups. At the same time, the richness of data formats and of the information that they carry make them useful evidence for a large variety of inquiries, and makes it easier to interrogate their reliability and quality in relation to different research conditions and aims. Phenomics data can potentially be used to answer many research questions. Cleaning by clustering in this case is not a cumulative process: it is crucial for researchers to lose as few data and metadata as possible, as one never knows what will turn out to be important later.

It has been frequently observed that big data aggregation is often accompanied by loss of contextual information (metadata).[16] While in both of our cases the role and ordering of contextual information plays a key role in the process of cleaning by clustering, the principles associated to handling such contextual information are considerably different. In economics, metadata become increasingly less relevant: the principles guiding data ordering and clustering are those of Prägnanz. In plant phenomics, metadata never cease to be relevant, as the principles guiding ordering and clustering are those of accuracy and traceability.

[15] Facts about phenomena, in the sense of Bogen and Woodward 1988.

[16] Lawrence Busch (2014, also discussed in Mittlestand and Floridi 2016) lists several reasons for this, including: Lossiness (lose aspects of the phenomena studied); Drift (phenomena change over time, but data representing them do not); Distancing (distance from phenomenon facilitates identification of patterns); Layering (reducing phenomena to set of variables, e.g. in Tidy data); Errors; Standards; Disproportionality; Amplification/reduction; Narratives.

Assumptions made about the nature of the phenomena at hand (respectively, plant morphology and business cycles) may seem to have a significant impact on the type of techniques and principles enacted by researchers. For instance, the proponents of MIAPPE explicitly note that

> we are fully aware that MIAPPE suggests a description of the experiment that is rather extended in comparison to current practices. Hence, although we think that all of the attributes in Table 1 are needed to adequately describe each dataset, we accept that, in practice, the full complement of information may not be possible to collect, or might be unavailable to the person building the dataset. Therefore, we have selected and marked those descriptors deemed absolutely essential. (Ćwiek-Kupczyńska et al. 2016, 7)

Remarkably, their "absolutely essential" list of traits still comprises 35 attributes, a skinnier list than the original list of over 80 attributes (ranging from 70 to over a hundred depending on growth conditions and type of environment/soil), but still daunting in its richness.

We do not think that these differences should be viewed simply as a measure of the difference between studying plants and studying economic conditions. Both types of phenomena are highly complex in their own ways, and arguably economic behaviour is even more difficult to reduce to a simple set of variables. A more plausible explanation lies in the methods and commitments characterizing the two fields of inquiry. Economics, business cycle analysis in particular, is a highly generalist field but it is not holistic: research focuses on analysing the business cycle as an isolated phenomenon. By contrast, plant phenomics favours a holistic approach, emphasising the complexity of the interrelated processes through which plant morphology is constituted (see Fig. 4 and also Leonelli 2016, ch. 6).

Furthermore, plant phenomics has no pretension to achieve a "complete representation" (or complete knowledge) of the plant systems it analyses, precisely because of their daunting complexity and the fact that so little is as yet known about them. Thus, any model proposed in plant science to analyse a phenomenon will be limited in scope, and need to be complemented by several others to provide a more comprehensive picture of the phenomena for specific investigative goals. Related to this, mathematical and statistical modelling – while of course strongly present in this work – are not always the primary or main tool of analysis; and their role is not always one of data validation, they are also employed as tools to order and display the data at hand in ways that may help analysis (Leonelli 2019).

5 Conclusions

Our analysis points to the difficulties experienced by analysts in providing general principles of cleanliness with regard to research data. This is nicely exemplified when considering the ongoing debate around the identification and application of overarching "tidy data principles" in contemporary data science, which seeks to outline criteria for "cleaning" and structuring data so as to make them amenable to computational analysis (Wickham 2014). Within this framework, data processing is

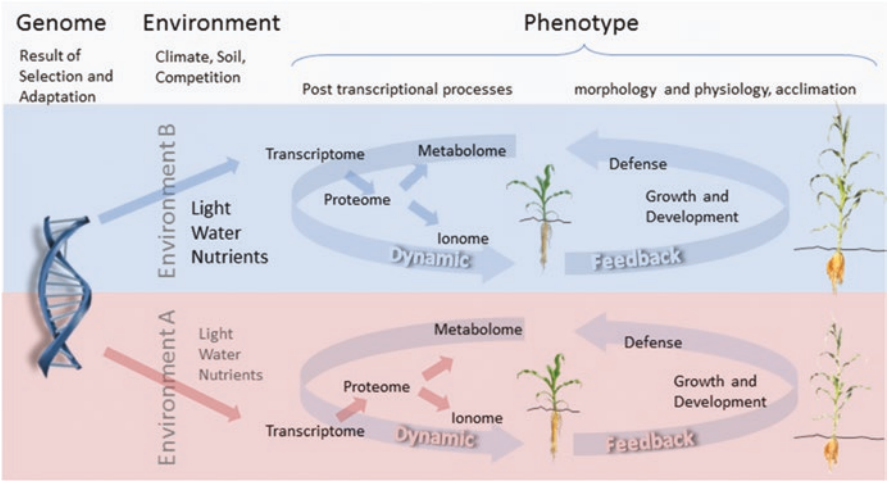


Fig. 4 Representation of the conceptual landscape for phenomics, taken from a seminal review paper from Walter et al. (2015)

conceptualised as consisting of four stages: (1) import data; (2) tidy data; (3) transform/visualise/model data; (4) communicate data. Tidy datasets are defined as providing "a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning)" (Wickham 2014, 2), thus helping to prepare data for visualisation and modelling. This literature does not shy away from data diversity, and recognises that data "tidiness" comes in a variety of different flavours depending on the field and goals of inquiry, the statistical and computational tools available (which are referred to as "tidy tools", p. 20), and the cognitive preferences of investigators. The starting point for this work is to acknowledge that determining what are observations and what are variables is relatively easy in the case of specific datasets, but that such a distinction is hard to define in general terms, also because of the diversity often characterising data sources and levels of abstraction. At the same time, an attempt is made to discuss tools through which "messy data" can be "tidied up", so as to be ready for computational analysis. An example is the activity of "melting", which consists of stacking datasets by turning columns of numbers into rows. Another is "string splitting", which involves splitting the columns of any given data table into different variables. Furthermore, a series of "tidy tools" are presented, such as data aggregation, filtering, visualisation and statistical modelling, whose common aim is to "take untidy datasets as input and return tidy datasets as outputs" (p. 12). All these strategies for cleanliness are meant to "make analysis easier by easing the transitions between manipulation, visualisation and modelling" (p. 15).

This approach to data cleaning aligns nicely with the strategy that we have called "cleaning by clustering". At the same time, our reading of Douglas's work on dirt provides a conceptual framework and rationale for this approach. It makes it clear that cleanliness is not a matter of removing unnecessary items, "noise" or "mess"

from somehow predefined "meaningful datasets", thus assuming that (1) there is a "best way" to order data regardless of the research aims of specific investigations; and (2) what researchers should consider as reliable and veritable data need to be uncovered and separated from "meaningless noise". By contrast, we propose to view data cleanliness as a process of ordering data into clusters, which runs in parallel with situated attempts to assign meaning to data in relation to specific research questions and goals. Thus cleaning can take a variety of different forms – and result in very different ideas of "what counts as data" – depending on the assumptions, commitments and circumstances of the research projects at hand. Moreover, our cases have shown that the above mentioned four stages of data analysis are actually four aspects of one process of data interpretation which cannot be separated from each other.

References

Bogen, James, and James Woodward. 1988. Saving the Phenomena. *Philosophical Review* 97 (3): 303–352.

Boumans, Marcel. 2015. *Science Outside the Laboratory*. Oxford: Oxford University Press.

Burns, Arthur F., and Wesley C. Mitchell. 1946. *Measuring Business Cycles*. New York: National Bureau of Economic Research.

Busch, Lawrence. 2014. Big Data, Big Questions | A Dozen Ways to Get Lost in Translation: Inherent Challenges in Large Scale Data Sets. *International Journal of Communication* 8 (0): 18. https://doi.org/10.1007/SpringerReference_22340.

Ćwiek-Kupczyńska, Hanna, Thomas Altmann, Daniel Arend, Elizabeth Arnaud, Dijun Chen, Guillaume Cornut, Fabio Fiorani, et al. 2016. Measures for Interoperability of Phenotypic Data: Minimum Information Requirements and Formatting. *Plant Methods* 12 (1): Bio Med Central: 44. https://doi.org/10.1186/s13007-016-0144-4.

Douglas, Mary. 2002[1966]. *Purity and Danger. An Analysis of the Concept of Pollution and Taboo*. London/New York: Routledge.

Halfmann, Gregor. this volume. Material Origins of a Data Journey in Ocean Science: How Sampling and Scaffolding Shape Data Practices. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Hoeppe, Götz. this volume. Sharing Data, Repairing Practices: On the Reflexivity of Astronomical Data Journeys. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Karaca, Koray. this volume. What Data Get to Travel in High Energy Physics? The Construction of Data at the Large Hadron Collider. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Leonelli, Sabina. 2011. Packaging Small Facts for Re-Use: Databases in Model Organism Biology. In *How Well Do Facts Travel?* ed. P. Howlett and M.S. Morgan, 325–348. Cambridge: Cambridge University Press.

———. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago: University of Chicago Press.

———. 2018. The Time of Data: Time-Scales of Data Use in the Life Sciences. *Philosophy of Science* 85 (5): 741–754.

———. 2019. What Distinguishes Data from Models? *European Journal for the Philosophy of Science* 9: 22.

———. this volume. Learning from Data Journeys. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Morgan, Mary S. 1990. *The History of Econometric Ideas*. Cambridge, MA: Cambridge University Press.

Palmer, Stephen E. 1999. *Vision Science*. Cambridge, MA: MIT Press.
Rogers, Susan, and Alberto Cambrosio. 2007. Making a New Technology Work: The Standardization and Regulation of Microarrays. *Journal of Biology* 80: 165–178.
Walter, Achim, Frank Liebisch, and Andreas Hund. 2015. Plant Phenotyping: From Bean Weighing to Image Analysis. *Plant Methods* 11 (1): 14. https://doi.org/10.1186/s13007-015-0056-8.
Wickham, Hadley. 2014. Tidy Data. *Journal of Statistical Software* 59 (10). https://doi.org/10.18637/jss.v059.i10.
Xavier, Alencar, Benjamin Hall, Anthony A. Hearst, Keith A. Cherkauer, and Katy M. Rainey. 2017. Genetic Architecture of Phenomic-Enabled Canopy Coverage in *Glycine Max*. *Genetics* 206 (2): 1081–1089. https://doi.org/10.1534/genetics.116.198713.

Marcel Boumans is Pierson Professor of History of Economics at Utrecht University. His main research focus is on understanding empirical research practices in social science from a combined historical and philosophy perspective. He is particularly interested in the practices of measurement and modelling and the role of mathematics in social science. Because models are not complete as sources of knowledge for sciences outside the laboratory, additional expert judgements are needed. This is the topic of his most recent monograph *Science Outside the Laboratory* (OUP, 2015). His current research project "Vision and Visualisation" focuses on exploring how expert judgments (views) are made and how they could be validated, particularly in those research practices where visualizations are made or used. The first outcome of this project is "Graph-Based Inductive Reasoning" published in *Studies in History and Philosophy of Science* (2016).

Sabina Leonelli is Professor of Philosophy and History of Science at the University of Exeter, where she codirects the Exeter Centre for the Study of the Life Sciences (Egenis) and leads the "Data Governance, Algorithms and Values" strand of the Institute for Data Science and Artificial Intelligence. Her research concerns the epistemology and governance of data-intensive science, the philosophy and history of organisms as scientific models and the role of open science in the global research landscape. She has an interest in science policy and served as expert for national and international bodies including the European Commission. She is a Turing Fellow, Editor-in-Chief of *History and Philosophy of the Life Sciences* and Associate Editor of the *Harvard Data Science Review.* Her publications span philosophy, social science, biology, history, data science and science policy and include the monographs *Data-Centric Biology: A Philosophical Study* (2016) and *La Recherche Scientifique à l'Ère des Big Data* (2019). Between 2014 and 2019, she led the European Research Council Starting Grant "The Epistemology of Data-Intensive Science" which supported the development of this volume.